Text-Prompted Diffusion Null-Space Model for Zero-Shot Image Restoration

Haijing Zhang (haijing@stanford.edu) Yining Mao (yiningm@stanford.edu)



Colorization



Inpainting



Deblurring

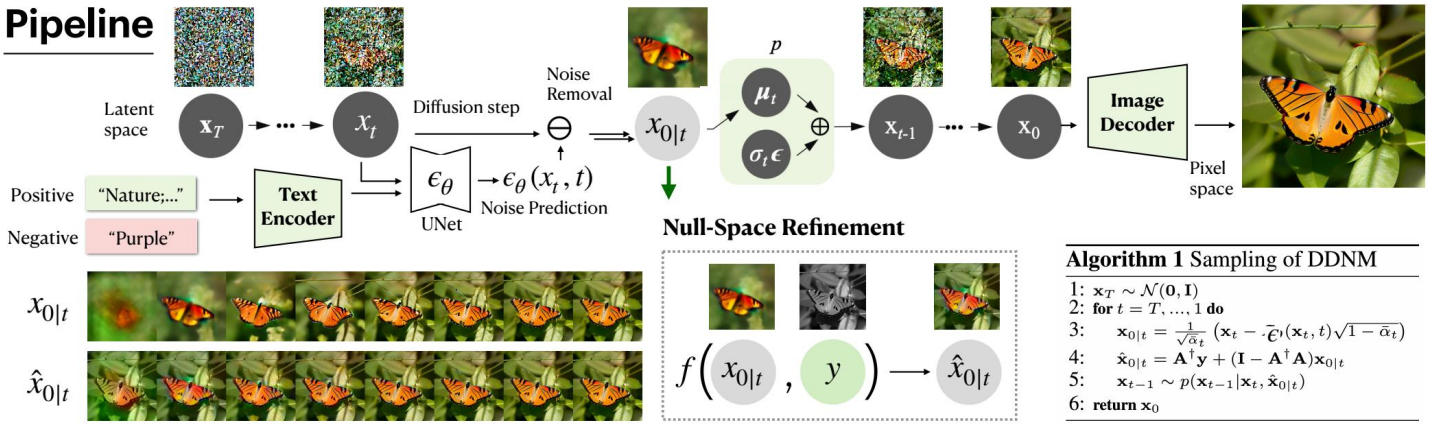
Abstract

This project extends the usage of the **Denoising Diffusion Null-Space Model (DDNM)** [1], the state-of-the-art zero-shot image restoration method. With a pre-trained diffusion model, DDNM is capable of producing **realistic and data-consistent results without any training** or modifications to the network structure. However, undesirable results are generated occasionally due to the inherent randomness in the diffusion models and the performance limitations related to the pretrained model and training dataset.

To combat this problem, we propose integrating DDNM with the **text-prompt-enabled latent diffusion model** [2], widely known as Stable Diffusion. We use text prompts to fix the undesirable components in the result image, making the restoration process more **flexible and robust**. By reimplementing both algorithms and conducting experiments on three tasks, we have demonstrated that DDNM, originally developed for pixel space, is compatible with latent space diffusion models, and that using text prompts can effectively and efficiently make adjustments to the restoration results.

[1] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model, 2022.
[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pages 10684–10695, 2022

Pipeline



Algorithm 1 Sampling of DDNM

$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, \ldots, 1$ **do**
  $\mathbf{x}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \bar{\epsilon}_\theta(\mathbf{x}_t, t)\sqrt{1-\bar{\alpha}_t}\right)$
  $\hat{\mathbf{x}}_{0|t} = \mathbf{A}^\dagger \mathbf{y} + (\mathbf{I} - \mathbf{A}^\dagger \mathbf{A})\mathbf{x}_{0|t}$
  $\mathbf{x}_{t-1} \sim p(\mathbf{x}_{t-1}|\mathbf{x}_t, \hat{\mathbf{x}}_{0|t})$
return $\mathbf{x}_0$

Denoising Diffusion Null-Space Model (DDNM)

Noise-free Image Restoration (IR)

Degraded image $y = Ax$ Ground-truth (GT) image

Linear degradation operator

Restore? $\hat{x}$

Constraints

Consistency $A\hat{x} \equiv y$ *Realness* $\hat{x} \sim q(x)$ Distribution of GT

Range-Null Space Decomposition

Range space Null space

$$x \equiv A^\dagger Ax + (I - A^\dagger A)x$$

General solution

$$\hat{x} \equiv A^\dagger y + (I - A^\dagger A)\bar{x}$$

$\bar{x}$: Predicted GT image

Multiply A on both sides

$$A\hat{x} \equiv y$$

This formula shows that no matter what $\bar{x}$ is, it would always satisfy the data consistency. The quality of $\bar{x}$ controls the realness.

For colorization:
A color → grey
$A^\dagger$ grey → color

Related Work

Traditional Model-based Method
Pros: less artifacts; Cons: fail in realistic details

$$\hat{x} = \arg\min_x \frac{1}{2\sigma^2}\|Ax - y\|_2^2 + \lambda R(x)$$

Data-fidelity Image-prior

End-to-End Deep Learning Method
Pros: high-quality; Cons:task specific; lack interpretability

$$\hat{x} = \arg\min_\theta \sum_{i=1}^{N} \|D_\theta(y_i) - x_i\|_2^2$$

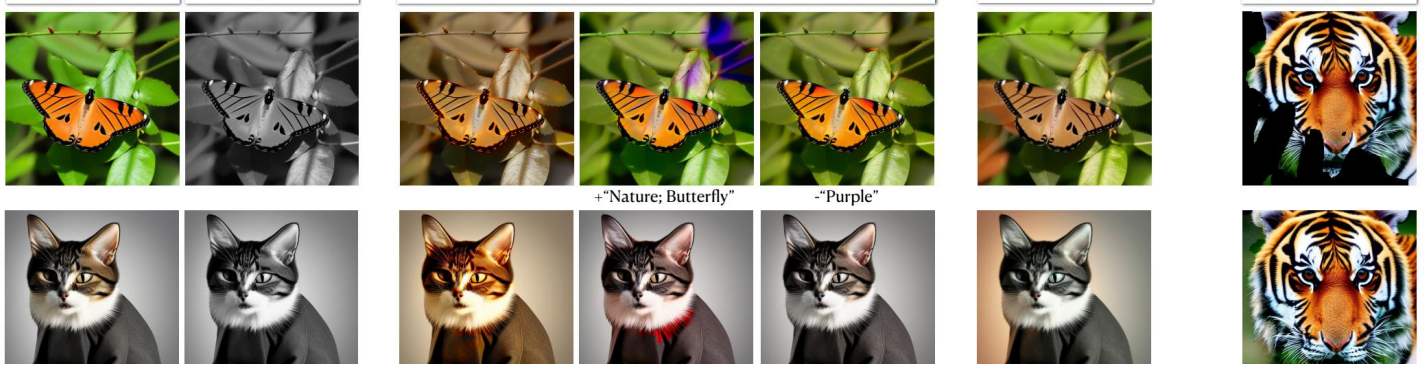
Deep Neural Network

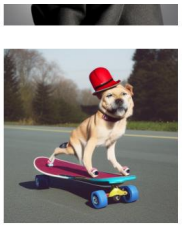
Zero-shot Pretrained Generative Model
Pros: good interpretability, efficient; Cons: hard to balance data consistency and realness

$$\hat{x} = \arg\min_w \frac{1}{2\sigma^2}\|AG(w) - y\|_2^2 + \lambda R(w)$$

Generative Model; Latent code

Experiments



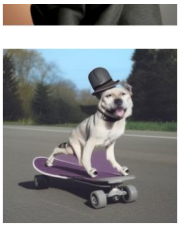


-“Yellow”

-“Red”

-“Yellow”

-“Red; Blue”



“ ”

+“Tiger...; background...”